doi:10.1016/j.jmb.2006.04.033 J. Mol. Biol. (2006) 359, 1456–1467

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Protein Refolding *in Silico* with Atom-based Statistical Potentials and Conformational Search Using a Simple Genetic Algorithm

Qiaojun Fang and David Shortle*

Department of Biological Chemistry, The Johns Hopkins University School of Medicine 725 N. Wolfe Street, Baltimore MD 21205, USA

A distance-dependent atom-pair potential that treats long range and local interactions separately has been developed and optimized to distinguish native protein structures from sets of incorrect or decoy structures. Atoms are divided into 30 types based on chemical properties and relative position in the amino acid side-chains. Several parameters affecting the calculation and evaluation of this statistical potential, such as the reference state, the bin width, cutoff distances between pairs, and the number of residues separating the atom pairs, are adjusted to achieve the best discrimination. The native structure has the lowest energy for 39 of the 40 sets of original ROSETTA decoys (1000 structures per set) and 23 of the 25 improved decoys (~1900 structures per set). Combined with the orientation-dependent backbone hydrogen bonding potential used by ROSETTA and a statistical solvation potential based on the solvent exclusion model of Lazaridis & Karplus, this potential is used as a scoring function for conformational search based on a genetic algorithm method. After unfolding the native structure by changing every phi and psi angle by either ±3, ±5 or ±7 degrees, five small proteins can be efficiently refolded, in some cases to within 0.5 Å C^{α} distance matrix error (DME) to the native state. Although no significant correlation is found between the total energy and structural similarity to the native state, a surprisingly strong correlation exists between the radius of gyration and the DME for low energy structures.

Keywords: statistical potentials; atom-pair potential; decoy discrimination; implicit solvation; simple genetic algorithm

*Corresponding author

Introduction

Strategies to predict protein structure from amino acid sequence reduce to two principal components: an energy function that assigns a score to the "quality" of a conformation and a search strategy for generating new conformations of higher quality. The terms in which this score is cast depend on whether the energy function is derived from physics-based principles, usually known as molecular mechanics potentials, or is based on knowledge of other protein structures, referred to as knowledge-based, database, or statistical potentials.[1] The energy equations of molecular mechanics score conformations in the real units of enthalpy. Statistical potentials, on the other hand, are calculated from the frequencies of occurrence of structural features in high resolution protein structures and then interpreted as free energies upon invoking the Boltzmann hypothesis.[2,3] Because the scientific foundations of statistical potentials are still rather questionable, some authors prefer they be developed in the framework of Bayesian statistics and probabilities, rather than as approximations to the true free energy.[1]

To improve the quality of models developed by homology-based or other types of modeling, interaction energies between individual atoms in protein structures must be calculated. Significant progress in this regard has been reported recently when these interaction energies are calculated with the Lennard–Jones 6-12 equation.[4] This molecular mechanics energy term, which models both the repulsive and attractive components of interaction between neighboring atoms, is also extensively used in

Abbreviations used: DME, distance matrix error; RG, radius of gyration; GA, genetic algorithm.

E-mail address of the corresponding author: shortle@jhmi.edu

simulations of protein behavior by molecular dynamics programs. While its mathematical form was chosen somewhat arbitrarily to permit rapid calculation, the 6-12 potential is the unchallenged standard for computational studies of systems involving thousands of atoms.

For decades, statistical potentials have been extensively applied to modeling side-chain/side-chain interactions and dihedral angle energies.[3,5,6–9] Yet their development and application at the atomic level have received little attention. In three published reports,[10–12] 167 different atom types were defined, in effect labeling every atom type in each of the 20 amino acids as chemically distinct. In addition, these atom-pair potentials included all pairs of atoms within distance cutoffs of 14.0 Å to 20.0 Å. Distances between pairs were subdivided into either 0.5 Å or 1.0 Å bins. In an earlier report,[13] 40 atom types were defined, and distances out to 10.5 Å were subdivided into 0.5 Å bins, although only 180 protein structures were used to compile pair statistics. In all four reports, the new potentials demonstrated significant power in identifying correct protein structures when challenged with a large number of mis-folded or decoy structures.

If a statistical potential is to model the energetics of interactions between atom pairs, it must be calculated in a manner that can capture the distance dependence of the energy, or the force between atoms. Subdividing the distance between atoms in 1 Å bins may be too coarse-grained to accurately describe the free energy as a function of distance. Including interactions between atoms at distances beyond which chemical principles predict a significant interaction serves only to slow the computation and increase the level of noise. And defining too many atom types can be expected to reduce the number of some data points to a level that is too small to yield accurate values.

Recent work from our group has demonstrated that the ability of statistical potentials to capture the energetics of local side-chain/backbone interactions requires a fine-grained description of distances and dihedral angles and thus a large database of proteins.[14–16] One common observation from this work has been the importance of designing the statistical potentials to model a hypothetical exchange reaction, in which an ''average'' residue is replaced with a specific residue in an otherwise identical structure. Here we report the development of a new statistical potential for atom-pairs plus an implicit solvation energy, both of which are based on this exchange reaction. Their high information content is demonstrated by their power to discriminate between the native structure and high quality all-atom decoy structures. More importantly, we demonstrate the utility of these potentials for protein structure prediction by employing them to direct a simple but efficient conformational search downhill in energy from a diverse set of unfolded conformations to structures similar to the native state.

Results

Optimization of the atom-pair potential

By definition, statistical potentials assign favorable scores to protein-like structural features and unfavorable scores to features that are uncommon or absent from proteins. When scored at the level of individual atoms, it seems reasonable that protein-like details will be difficult to achieve for incorrect or so-called ''decoy'' structures. If the native structure is the conformation with the lowest real energy and a decoy has little in common with the native structure, it can be assumed that the decoy should score poorly with an energy function that approximates the true one. We make this the central working assumption here and optimize the adjustable parameters in these statistical potentials to maximize the energy gap between the native structure and the set of decoy structures.

The atom-pair potential is optimized by several rounds of recalculation followed by comparison of the score between native structures and their sets of ROSETTA decoys. Both the rank score of the wild-type structure relative to the members of the decoy set and the energy difference (energy gap) between wild-type and the lowest energy decoy are evaluated. The path to an optimum consisted of first finding a local optimum for parameter *A* and then evaluating or re-evaluating one or more different parameters employing this standard value for *A*. Thus, no claim can be made that the true optimum has been found.

Since in previously published atom potentials, the distance between atom pairs has been coarsely grained,[10–12] we first examine the effect of subdividing this distance into 0.1, 0.2, 0.4, 0.5, 0.8, and 1.0 Å bins. As can be seen in Table 1, similar results in decoy discrimination are obtained with 0.2, 0.4, and 0.5 Å bins, but the 0.2 Å gives slightly better performance as assessed by the average energy gap. Since an interval of 0.2 Å also seems adequate to model the energetics of atomic interactions, it is chosen as the standard for subsequent calculations.

The maximum distance between atom surfaces included in the energy calculation is varied from 3 Å to 10 Å. Not surprisingly, the results for all distances from 4 Å to 10 Å are quite similar (data not shown). Since the 4 Å cutoff performs slightly better, leads to faster evaluation of the energy, in principle, includes less noise, it is chosen as the standard value.

To eliminate the effects of structural constraints on the independence of atom-pair formation, the minimal number of residues that must intervene between two atoms treated as a pair is varied from 0 to 5. As shown in Table 2, two residues are not sufficient; rather three to five residues give better results, with four (i.e. *i* to *i* +5) being the value set as standard. The most effective use of the potential calculated with a minimum of four separating residues, however, requires that it be applied to almost all atom-atom interactions. Since the best result is obtained with one separating residue (Table